**INTERVIEW CASE ASSIGNMENT**

****

**Group - 6**

Bansi Vaishnav

Bargavi Kongara

Shailja Sharma

Srivalli Adigarla

BAN-5501 - DATA MGMT & SQL FOR ANALYTICS

**Prof. Kathy Fogel**

1. **Review unstructured JSON data and diagram a new structured relational data model.**

The unstructured data has three tables:

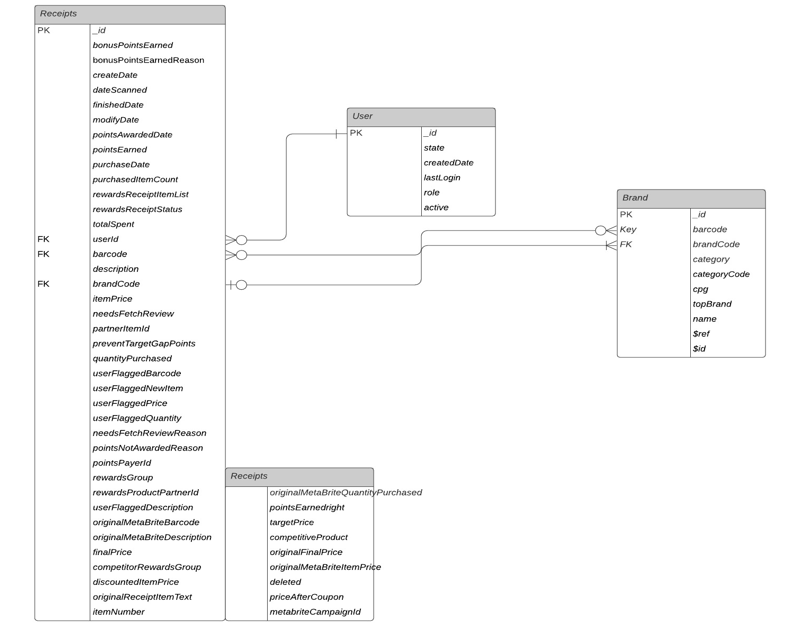
**Receipts:** This table stores information about each receipt, including the receipt ID, the  date      the receipt was issued, the bonus points earned for the purchase, and the user ID of the person who made the purchase.

**Brand:** This table stores information about each brand, name, cpg, including the brand ID and code.

**Users:** This table stores information about each user, including the user ID, state, last login information, role, etc.

* The Receipts table and the Brands table are connected via brand and bar codes, allowing us to track sales.
* The Users table and the Receipts table are connected via the user ID field, which allows us to track which user made each purchase.

**Relational data model:**



1. **Generate a query to capture data quality issues against the new structured relational data model.**

**Question:** When considering average spend from receipts with ‘rewardsReceiptStatus’ of ‘Accepted’ or ‘Rejected’, which is greater?

SELECT AVG(totalSpent)

FROM receipts

WHERE rewardsReceiptStatus=”Accepted”;

The above code gives the average spent on receipts with the rewards status “Accepted”.

SELECT AVG(totalSpent)

FROM Receipts

WHERE rewardsReceiptStatus=”Rejected”;

This gives the average spend on receipts with rewards status “Rejected”.

We compare the values returned by the above statements to find the greater among the two.

1. **Generate a query to capture data quality issues against the new structured relational data model.**

While working with data, we always checked for invalidity and incompleteness. We tested the values for a few fields using GROUP BY SQL operations to see whether they matched the expectations. Sadly, there were numerous rows with either inaccurate or missing values.

Some items from the Receipt table had either a missing barcode or an ITEM NOT FOUND description.

There were also redundancies with the various price columns from the Receipt table.

1. **Write a short email or Slack message to the business stakeholder.**

**Subject**: Data quality issues in Receipts, users and Brands records

Dear Kevin,

I hope this email finds you well. I’m wiring this email to let you know few of our findings regarding data quality issues in Receipts, Users and Brands records after conducting an exploratory analysis. Below are the findings:

* There's a considerable amount of missing data for fields like finishedDate, pointsEarned, purchasedItemCount, totalSpent, rewardsReceiptItemList, topBrand, and categoryCode.
* Some values in the pointsEarned, purchasedItemCount, and totalSpent fields appear to be out of place and should be investigated.
* More than half of the User’s records have duplicates that should be eliminated.
* Date formats are inconsistent and should be reviewed for consistency.

I have a plan to deal with the missing values issue and other problems as well, that I’d like to discuss with you in detail. Let me know what time works best for you so that we can set up a meeting.